STRUCTURE-INDEPENDENT VISUAL MOTION CONTROL ON THE ESSENTIAL MANIFOLD

S. SOATTO* and P. PERONA * **

**Control and Dynamical Systems (CDS) – California Institute of Technology 116-81, Pasadena–CA 91125. E-mail soatto@systems.caltech.edu*

***Università di Padova, Dipartimento di Elettronica ed Informatica. Via Gradenigo 6/A 35100 Padova–Italy.*

Abstract. A new framework for visual motion control is described, which consists of formulating the control task on the so-called *essential manifold*, a "compact" matrix representation of $SE(3)$. Unlike previous image plane control techniques, our method does not require information about the geometric structure (depth) of the scene or target. This allows us to design control laws that are not ill-conditioned when close to zero-disparity configurations. The control relies on a causal motion estimator (called the "essential filter") that identifies recursively an implicit dynamical model with parameters on the essential manifold.

Key Words. Visual motion estimation, essential manifold, motion control

1. INTRODUCTION

For humans to perceive the geometric structure and motion of objects within their environment and accomplish tasks such as navigating, grasping and manipulating them, vision is of vital importance. Recently, the improvements in hardware technology and the application of tools from the theory of nonlinear control and estimation (see for example (Ghosh *et al.*, 1994)) have let to encouraging results in the application of computer vision to autonomous navigation (Dickmanns and Graefe, 1988), tracking (Lei and Ghosh, 1993; Chaumette and Santos, 1993), manipulation (Ebine and Kimura, 1991), docking (Ho and McClamroch, 1992) and planning (Curwen *et al.*, 1992).

In an artificial context, perceiving the environment corresponds to estimating the motion and the structure of each object that is visible with a camera. In the past, a number of techniques have been proposed to solve the above problem for a single rigid object (scene) represented by a small number of "feature points" in 3-D space. Those points correspond to the position in space of "salient features" of the image plane, such as regions with high spatial brightness gradient.

We assume that a number of feature points is available as well as the correspondence of their projections across time[1]. Since we aim at real time applications, we are interested only in recursive and causal schemes that do not make *a priori* assumptions about the structure of the scene. There are few such schemes available in the literature (Azarbayejani *et al.*, New York, 1993; Heel, March 1991), which are intrinsically local and based upon an Extended Kalman Filter (EKF) (Kalman, 1960; Jazwinski, 1970).

In this paper we investigate the use of a dynamic visual motion estimator in the feedback loop of a motion control system. In section 4 we formulate the control problem on a differentiable manifold, called the essential manifold. We will see that the essential manifold, as an alternative and "compact" representation of $SE(3)$, is particularly well suited for applications involving vision and motion control. In the experimental section 5 we test a simple control strategy on the essential manifold and contrast traditional controllers based on a formulation of the task on the image plane.

In order to make the paper self contained, we describe some of the properties of the essential manifold in section 2. Here we also introduce a characterization of the essential manifold as the tangent bundle of $SO(3)$. In section 3 we review the principles of a recursive motion estimation scheme on the essential manifold which was introduced in (Soatto *et al.*, May 1994*a*).

2. REPRESENTATION OF RIGID MOTION VIA THE ESSENTIAL MANIFOLD

The movement of a rigid body in $\mathbb{R}^3$ can be described by a point in the Euclidean group of transformations, $g(t) \equiv (T(t), R(t)) \in SE(3)$, which acts on points of

[1] There are a number of methods for performing feature tracking/optical flow as well as for selecting automatically "good features"; some are described and compared in (Barron *et al.*, 1992).

$\mathbb{R}^3$ via[2]

$$\mathbf{X}(t+1) = R(t)(\mathbf{X}(t) - T(t)). \tag{1}$$

We measure the *projection* of each feature point $\mathbf{X}^i$, $\forall i = 1 \dots N$ on the image plane:

$$\begin{aligned} \pi : \mathbb{R}^3 &\rightarrow \mathbb{RP}^2 \\ \mathbf{X}^i &\mapsto \mathbf{x}^i = [\frac{\mathbf{X}_1^i}{\mathbf{X}_3^i} \ \frac{\mathbf{X}_2^i}{\mathbf{X}_3^i} \ 1]^T \end{aligned} \tag{2}$$

where the last expression describes an ideal perspective projection with unitary focal length. The space $SE(3)$ can be embedded in the matrix group $\mathcal{GL}(4)$, and the matrix product used as group operation, via the homogeneous coordinates. A more "compact" representation of a rigid motion (T, R) can be derived from the so-called "essential matrices", which are elements of the subset of $\mathbb{R}^{3\times3}$

$$\tilde{E} \doteq \{RS \mid R \in SO(3) \ S \doteq T\wedge \in so(3)\}. \tag{3}$$

The essential space $\tilde{E}$ was introduced in (Longuet-Higgins, 1981) and, since then, it was shown to have the structure of an algebraic variety (Maybank, 1992) and of a topological manifold (Soatto *et al.*, 1994*b*).

2.1. *ALGEBRAIC STRUCTURE OF* $\tilde{\mathbf{E}}$

It is proven (Demazure, see (Faugeras, 1993)) the space of essential matrices is the algebraic variety characterized as the subset of the 3×3 matrices that satisfy the following polynomial equations:

$$\mathbf{Q} \in \tilde{E} \subset \mathbb{R}^{3\times3} \Leftrightarrow \begin{cases} \det(\mathbf{Q}) = 0 \\ \frac{1}{2}\mathrm{tr}(\mathbf{Q}\mathbf{Q}^T)\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T\mathbf{Q}. \end{cases} \tag{4}$$

This characterization leads to algebraic methods for estimating essential matrices from pairs of images, which we are not considering here due to their sensitivity to the image noise. The interested reader may consult (Longuet-Higgins, 1981; Faugeras, 1993; Maybank, 1992).

There are also some relationships between the essential manifold and the space of "dual quaternions"[3] (see for example (Chevallier, 1991) and references therein).

2.2. *DIFFERENTIAL STRUCTURE OF* $\tilde{\mathbf{E}}$

The essential manifold can also be characterized as the tangent bundle of $SO(3)$. Consider $(T, R) \in SE(3)$, then $S \doteq T\wedge \in so(3) = T_eSO(3)$ is a tangent vector to the origin of $SO(3)$, which can be push-forwarded by left translation at any location R of $SO(3)$. Therefore the tangent vector to $SO(3)$ at R in the direction of S is given by $\mathbf{Q} = RS \in T_RSO(3)$ which is an essential matrix:

$$\mathbf{Q} \in \tilde{E} \Leftrightarrow \mathbf{Q} \in TSO(3). \tag{5}$$

From this we can convince ourselves that the essential space is a differentiable manifold of dimension six.

2.3. *TOPOLOGICAL STRUCTURE OF* $\mathbf{E}$

In order to be able to formulate a motion estimator on the essential manifold, we want to find a homeomorphism between the essential manifold and the Euclidean space $\mathbb{R}^6$. Since there is a structural scale ambiguity when recovering motion from visual information (Longuet-Higgins, 1981), we restrict our attention to the set of essential matrices of unitary 2-norm, $E \doteq \{\mathbf{Q}/\|\mathbf{Q}\|_2 \mid \mathbf{Q} \in \tilde{E}\}$, which is a five-dimensional manifold[4]. Consider the map

$$\begin{aligned} \Phi : E &\rightarrow S^2 \times \mathbb{R}^3 \\ \mathbf{Q} = U\Sigma W^T &\mapsto (V, \Omega) \end{aligned} \tag{6}$$

where U, Σ, W are determined by the Singular Value Decomposition (SVD) of $\mathbf{Q}$, and (V, Ω) are the canonical (exponential) coordinates (Murray *et al.*, 1993) of $(T, R) \doteq (\pm W_{\cdot 3}, UR_{X_3}(\pm\frac{\pi}{2})W^T)$. $W_{\cdot 3}$ is the third column of W and $R_{X_3}(\theta)$ is a rotation of θ about the axis X_3. The inverse function is simply $\Phi^{-1}(V, \Omega) = R(T\wedge)$, where (T, R) correspond to (V, Ω) via the exponential parametrization. Note that there are two sign ambiguities in the definition of the local coordinates of the "normalized" essential manifold. Those ambiguities can be resolved in the context of visual motion estimation by imposing the "positive depth constraint" (Longuet-Higgins, 1981), as described in (Soatto *et al.*, May 1994*a*). We can choose the correct sign combination by checking that the visible points give a reconstructed depth which is *positive* (i.e. they lie in front of the viewer). Once this choice is made at the first time instant, it can be propagated across time, so that the structure of the local coordinates automatically imposes positive depth.

The essential matrices may also be characterized as the 3×3 matrices having two equal singular values and a zero singular value (Huang & Faugeras, see (Faugeras, 1993)):

$$\mathbf{Q} = U\Sigma W^T \in \tilde{E} \Leftrightarrow \Sigma = \mathrm{diag}(\sigma, \sigma, 0) \tag{7}$$

where $\sigma \in \mathbb{R}^+$. The above result can be used in order to define a "projection" of an arbitrary 3×3 matrix M with SVD $\tilde{U}\tilde{\Sigma}\tilde{W}^T$ onto the normalized essential manifold E which has minimum Frobenius and 2-norm properties:

$$\mathrm{pr}_E(M) = \tilde{U}\mathrm{diag}(1, 1, 0)\tilde{W}^T. \tag{8}$$

Now that we have seen some of the structure of the essential manifold, we describe a scheme for estimating a rigid motion, interpreted as a point on the essential manifold, from visual information.

---

[2] Here $R(t)$ represents the orientation of the reference at time t relative to the reference at time $t+1$, while $T(t)$ is the position of the origin of the reference at time $t+1$ in the reference of time t. Such a notation is chosen for consistency with the standard notation of the essential matrices (Longuet-Higgins, 1981).

[3] The relation between the essential manifold and the space of dual quaternions was suggested to us by J. Burdick.

[4] We will be able to estimate the direction of translation, but not its norm, unless some scale information is available about the scene at some time instant.

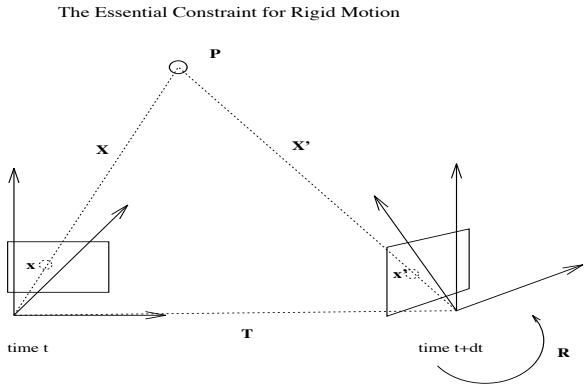

Fig. 1. *The coplanarity constraint.*

3. VISUAL MOTION ESTIMATION ON THE ESSENTIAL MANIFOLD

In this section we describe a recursive scheme for estimating the rigid motion of an object viewed under perspective projection. The scheme is based upon the identification of a nonlinear and implicit dynamical model, with parameters on the essential manifold. Such a model exhibits global as well as local observability/identifiability properties, and is independent of the structure of the scene (i.e. of the depth of the feature points). The identification can be carried out in the local coordinates of the essential manifold, or in its embedding space. The resulting schemes differ by the statistical model of motion employed: the first assumes a model of motion as a random walk in $\mathbb{R}^5$ lifted to the essential manifold, and solves a nonlinear estimation problem on a linear space. The second assumes a model of motion as a random walk on $\mathbb{R}^9$ "projected" onto the essential manifold, and solves at each step a *linear* update on the embedding space followed by a projection of the estimate onto the manifold. In the following subsections we give a heuristic derivation of the filter. For a detailed description, the reader may consult (Soatto *et al.*, 1994*b*; Soatto, 1994).

3.1. *THE ESSENTIAL FILTER IN LOCAL COORDINATES*

Consider a rigid motion between two time instants. Given any transformed point $\mathbf{X}^i$, its coordinates in the reference frame at time t, the corresponding coordinates in the reference at $t+1$ and the translation vector are coplanar (see figure 1). The same holds for $\mathbf{x}^i$ in place of $\mathbf{X}^i$, since the two are parallel. For each visible point we can write the coplanarity constraint, for example in the reference frame at time t, as

$$\mathbf{x}^i(t+1)^T R(t) \left(T(t) \wedge \mathbf{x}^i(t)\right) = 0 \qquad \forall i. \tag{9}$$

We measure directly the image plane coordinates $\mathbf{x}^i \in \mathbb{RP}^2$ up to some noise, which we safely assume to be white, zero-mean and Gaussian:

$$y^i(t) = \mathbf{x}^i(t) + n^i(t) \qquad n^i \in \mathcal{N}(0, R_{n^i}) \tag{10}$$

The estimation of motion amounts therefore to identifying the following implicit dynamical model with parameters on the essential manifold

$$\begin{cases} \mathbf{x}^i(t+1)^T \mathbf{Q}(t) \mathbf{x}^i(t) = 0 & \mathbf{Q} \in E \\ y^i(t) = \mathbf{x}^i(t) + n^i(t) & \forall i = 1 \dots N. \end{cases} \tag{11}$$

If we apply to the previous system the local coordinates homeomorphism Φ defined in eq. (6), we can write a corresponding estimation model in the local coordinates $\mathbb{R}^5$: let $\xi \doteq (V, \Omega)$, then

$$\begin{cases} \xi(t+1) = \xi(t) + n_\xi(t) & \xi \in \mathbb{R}^5 \\ y^i(t)^T \Phi^{-1}(\xi(t)) y^i(t-1) = \tilde{n}_i(t) & \forall i \end{cases} \tag{12}$$

where $n_\xi(t)$ is the white noise that drives the random walk model, and $\tilde{n}^i(t)$ is an induced residual noise whose second order statistics can be characterized in terms of the variance of the measurement error $n^i(t)$ (Soatto *et al.*, 1994*b*).

Note that if a dynamical model for motion is available, as for example when the camera is mounted on a moving vehicle or on a robot arm, we can substitute the random walk model with a dynamical model of the form $\xi^i(t+1) = f^i(\xi(t), n_\xi(t))$, where now n_ξ describes the state of the vehicle or robot arm.

The state of the model of eq. (12) is defined on a linear space and can now be estimated using a variation of the Extended Kalman Filter for implicit measurement constraints, which is derived in (Soatto *et al.*, 1994*b*). We summarize here the equations of the estimator. Write the coplanarity constraint (11) for N points in the form of a matrix equation $\chi \mathbf{Q} = 0$, where χ is a $N \times 9$ matrix and $\mathbf{Q} = \Phi^{-1}(\xi)$ is intended as a nine-dimensional column vector. Call $C \doteq \left(\frac{\partial \chi \Phi^{-1}}{\partial \xi}\right)$ and $D \doteq \left(\frac{\partial \chi \Phi^{-1}}{\partial \mathbf{X}}\right)$, R_α the variance of the process α, then we have

Prediction step:

$$\begin{aligned} \hat{\xi}(t+1|t) &= \hat{\xi}(t|t) \; ; \; \hat{\xi}(0|0) = \xi_0 \\ P(t+1|t) &= P(t|t) + R_\xi \; ; \; P(0|0) = P_0 \end{aligned}$$

Update step:

$$\begin{aligned} \hat{\xi}(t+1|t+1) &= \hat{\xi}(t+1|t) - \\ &\quad - L(t+1)\chi(t+1)\Phi^{-1}(\hat{\xi}(t+1|t)) \\ P(t+1|t+1) &= \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + \\ &\quad + L(t+1)R_{\tilde{n}}(t+1)L^T(t+1) \end{aligned}$$

Gain:

$$\begin{aligned} L(t+1) &= P(t+1|t)C^T(t+1)\Lambda^{-1}(t+1) \\ \Lambda(t+1) &= C(t+1)P(t+1|t)C^T(t+1) + \\ &\quad + R_{\tilde{n}}(t+1) \\ \Gamma(t+1) &= I - L(t+1)C(t+1) \end{aligned}$$

Innovation variance:

$$R_{\tilde{n}}(t+1) = D(t+1)R_n D^T(t+1)$$

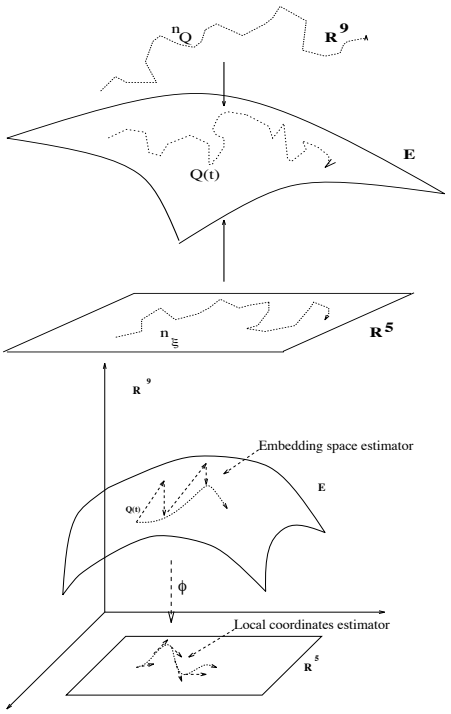

Fig. 2. *(Top) model of motion as a random walk in* $\mathbb{R}^5$ *lifted to the manifold or as a random walk in* $\mathbb{R}^9$ *projected onto the manifold. (Bottom) estimation on the Essential Space.*

3.2. *THE ESSENTIAL FILTER IN THE EMBEDDING SPACE*

The model (11) is *linear* in $\mathbf{Q}$. However, it is defined on a state-space which is *not a linear space*. We could thus think of lifting the model to the (linear) embedding space $\mathbb{R}^9$, and at each step "project" the current estimate onto the manifold. In general, this could be a very bad idea, for we perform the update in a bigger space, and then impose the structure of the state manifold *a posteriori*. In this case, however, we can show that the only difference between the filter in local coordinates and the filter defined in the embedding space is *the model of motion* employed: in the first case it is a random walk on $\mathbb{R}^5$ *lifted to the essential manifold*, whereas in the second case it is a random walk in $\mathbb{R}^9$ *projected onto the manifold*. Therefore we reduce the comparison between the two schemes to a *modeling issue*, which can be assessed only a posteriori (see figure 2).

We define the operation $\oplus$ as the projection onto the essential manifold of the sum of two essential matrices interpreted as elements of $\mathbb{R}^{3\times3}$: $\mathbf{Q}_1 \oplus \mathbf{Q}_2 \doteq \mathrm{pr}_E(\mathbf{Q}_1 + \mathbf{Q}_2)$. The filter defined in the embedding space is *linear*, and solves optimally (in the sense of least error variance) the prediction for the model

$$\begin{cases} \mathbf{Q}(t+1) = \mathbf{Q}(t) \oplus n_{\mathbf{Q}}(t) & \mathbf{Q}(t) \in E \\ y^i(t)^T \mathbf{Q}(t) y^i(t-1) = \tilde{n}^i(t) & \forall i = 1 \dots N \end{cases} \tag{13}$$

The solution of the estimation task is derived in (Soatto *et al.*, 1994*b*) and can be summarized as follows:

Prediction step:

$$\begin{aligned} \hat{\mathbf{Q}}(t+1|t) &= \hat{\mathbf{Q}}(t|t)\ ;\ \hat{\mathbf{Q}}(0|0) = \mathbf{Q}_0 \\ P(t+1|t) &= P(t|t) + R_{\mathbf{Q}}\ ;\ P(0|0) = P_0 \end{aligned}$$

Update step:

$$\begin{aligned} \hat{\mathbf{Q}}(t+1|t+1) &= \hat{\mathbf{Q}}(t+1|t) \oplus \\ &\oplus\ L(t+1)\chi(t+1)\hat{\mathbf{Q}}(t+1|t) \\ P(t+1|t+1) &= \Gamma(t+1)P(t+1|t)\Gamma^T(t+1) + \\ &+\ L(t+1)R_{\tilde{n}}(t+1)L^T(t+1) \end{aligned}$$

Gain:

$$\begin{aligned} L(t+1) &= -P(t+1|t)\chi(t+1)\Lambda^{-1}(t+1) \\ \Lambda(t+1) &= \chi(t+1)P(t+1|t)\chi(t+1) + \\ &+\ R_{\tilde{n}}(t+1) \\ \Gamma(t+1) &= I - L(t+1)\chi(t+1) \\ R_{\tilde{n}}(t+1) &= D(t+1)R_n D^T(t+1) \end{aligned}$$

4. VISUAL MOTION CONTROL

We are now ready to use the essential filter in the feedback loop of motion control systems. Traditionally, the control task in systems using vision as a sensor has been formulated directly on the image plane (Ebine and Kimura, 1991). This choice is very natural in certain applications, for example tracking, docking, navigation etc.. However, it results in methods that are intrinsically local, whereas there are applications in which one is required to track a globally prescribed path in the full configuration space. Furthermore, the control on the image plane exhibits some limitations due to the dependence of the controller on the structure (depth) of the observed scene.

In the next subsection we briefly describe a simple tracking control on the image plane, and highlight its limitations. In the following subsection we propose to formulate the tracking problem in the configuration space in its essential representation (through the essential manifold). We anticipate the resulting control has more "global" features and does not depend on the structure of the observed scene. Instead, structure comes as a byproduct of the essential estimator once the control task has been accomplished. In the experimental section we describe simulated experiments of the behavior of a simple controller based on the essential filter. We suppose the camera is mounted on a moving platform, on which we have full control. For simplicity we neglect the dynamic constraints and assume to be able to control directly the translational and rotational velocity of the platform. Suppose our task is to maintain a given relative configuration between the platform and the scene. Such situation occurs in tracking the motion of a three dimensional object (of unknown shape and kinematics) or in maintaining a fixed pose with respect to a scene despite the action of disturbances on the platform (as for example in hovering or in underwater operation).

4.1. *CONTROL ON THE IMAGE PLANE*

Consider the model defined by (1)(2), and the following expression for the time derivative of the output $\mathbf{x}$ (also called "motion field" or "real optical flow"):

$$\dot{\mathbf{x}}^i(t) = \mathcal{J}(\mathbf{x}^i(t), \mathbf{X}_3^i(t))u(t) \quad \forall i = 1 \dots N \tag{14}$$

where $\mathcal{J}(x, X_3) \doteq$

$$\begin{bmatrix} \frac{1}{X_3} & 0 & -\frac{x_1}{X_3} - x_1 x_2 & 1 + x_1^2 & -x_2 \\ 0 & \frac{1}{X_3} & -\frac{x_2}{X_3} - (1 + x_2^2) & x_1 x_2 & x_1 \end{bmatrix} \tag{15}$$

and $\mathbf{x}^i$ indicates the image plane coordinates of the projection, while $\mathbf{X}_3^i$ denotes the third component of the space coordinate (depth) of each point. The vector $u(t) \doteq (V(t), \Omega(t))$ is the canonical exponential representation of the instantaneous motion $(T(t), R(t))$. Suppose the initial configuration of the points on the image plane is $\mathbf{x}^i(t_0|t_0) = \mathbf{x}_0^i$, and an exogenous agent acts by moving either the platform on which the camera is mounted or the target which the camera is looking at, producing a deformation of its image:

$$\mathbf{x}^i(t+1) = \mathbf{x}^i(t) + \tilde{\mathbf{x}}^i(t). \tag{16}$$

Suppose our goal is to keep the configuration of the observed points fixed at the value of the initial instant $\mathbf{x}_0^i$. At any time we can measure a noisy version of the instantaneous configuration modified by the external agent, and act with the control of the platform on which the camera is mounted. Using a first step approximation, one could write

$$\begin{aligned} \mathbf{x}^i(t+1) \cong\ & \mathbf{x}^i(t) + \\ & + \mathcal{J}(\mathbf{x}^i(t), \hat{\mathbf{X}}_3^i(t))u(t) \end{aligned} \tag{17}$$

and use a minimum time controller:

$$\begin{aligned} u(t) \doteq\ & \mathcal{J}^\dagger(\mathbf{x}^i(t), \hat{\mathbf{X}}_3^i(t)) \cdot \\ & \cdot (x_0^i - \mathbf{x}^i(t)) \end{aligned} \tag{18}$$

where $\dagger$ denotes the pseudoinverse. Note that the control depends on the depth of each point of the scene $\hat{\mathbf{X}}_3^i(t)$. Such a strategy has been experimented by (Ebine and Kimura, 1991), who pioneered the control on the image plane. However, the expression of the deadbeat controller on the image plane depends on the inverse depth of each visible points, which needs to be "estimated" on line. This problem can be overcome by assuming that *the structure of the scene is known*, and therefore the inverse depth can be recovered linearly (the so-called "calibration" phase). Another alternative, which we do not pursue here, is the use of a stereo system.

If the structure (depth) of the scene is not known, we need to *estimate* it, unless the motion of the target is purely rotational about the center of the viewer's reference, in which case $\mathcal{J}$ does not depend on the depth. In order to estimate depth, we need non-zero *disparity* (also called visual parallax), which is the displacement of corresponding points across different images. When disparity is close to zero, the recovery of the depth is ill conditioned (Soatto *et al.*, 1993). Therefore the image based controller, which depends on the depth, tries to drive the system towards a configuration of zero disparity, which does not allow to recover depth. As a result the controller either "drifts" or "swings", as is discussed in the experimental section.

4.2. *CONTROL ON THE ESSENTIAL MANIFOLD*

Consider $\mathbf{Q}_0 \in E$ describing the relative configuration between the scene and the platform at the initial instance, and suppose we ask it to be constant despite the motion of the scene, encoded by an arbitrary $d(t) \in E$. We indicate with $\mathbf{Q}(t)$ the essential matrix describing the motion between the *initial* instant and the current time, which is therefore defined by the essential constraint $\mathbf{x}^i(t)^T \mathbf{Q}(t) \mathbf{x}_0^i \doteq 0$. Note that usually in the essential filter we consider $\mathbf{Q}(t)$ to be the instantaneous configuration with respect to the observer's reference at the previous time sample. The effect of the exogenous displacement (motion of the scene) and the control action are described by the model

$$\begin{cases} \mathbf{Q}(t+1) = \mathbf{Q}(t) \tilde{+} \Phi^{-1}(u(t)) \tilde{+} d(t) \\ y^i(t)^T \mathbf{Q}(t) y(0) = \tilde{n}(t) \end{cases} \tag{19}$$

where $\tilde{+}$ represents the sum of the local coordinates, $\tilde{n}$ describes the effect of the estimation error (it is in fact the pseudo-innovation of the essential filter). In general we may want to specify the control task in terms of some *distance* defined on the essential space, $d_E(\mathbf{Q}_1, \mathbf{Q}_2)$, so that

$$e(t) \doteq d_E(\mathbf{Q}(t), \mathbf{Q}_d(t)) \tag{20}$$

satisfies a difference equation whose dynamics can be assigned by choice of the input.

4.3. *CHOICE OF A METRIC ON THE ESSENTIAL MANIFOLD*

Since E can be interpreted as an alternative representation of $SE(3)$, any control strategy on the Euclidean group can be mapped onto the essential manifold. However, if we were able to formulate the control strategy directly on the essential manifold, the essential filter would then gives us a direct estimate of the full state which is optimal, independent of the structure and obtained linearly from the visual data (Soatto *et al.*, May 1994*a*).

The choice of a metric on the essential space is not a trivial issue, and we intend in this paper to hint at some possible choices. First of all any metric in the Euclidean space $SE(3)$ can be "mapped" onto the essential manifold by defining

$$d_E(\mathbf{Q}_1, \mathbf{Q}_2) \doteq d_{SE(3)}(\Psi^{-1} \circ \Phi(\mathbf{Q}_1), \Psi^{-1} \circ \Phi(\mathbf{Q}_2))$$

where Ψ and Φ are local coordinatizations of $SE(3)$ and E respectively. An alternative (and equivalent) method is to set the metric directly in the local coordinates and then "lift" it to the manifold. It must be pointed out, however, that there is no natural (invariant) choice of a metric on the Euclidean group. Another possibility is to "project" a metric of the am-

bient space of the essential manifold, $\mathbb{R}^9$, by using the projection onto the manifold pr_E. It is unclear at the moment what the properties of such a metric may be. Note also that a possible way of generating a path between two points of the essential manifold, based on its interpretation as the tangent bundle of $SO(3)$, is to formulate a control that connects two points of $SO(3)$ with a given direction in the tangent plane. Such control strategies, called "dynamic interpolation" have been studied for Riemannian manifolds and Lie groups by (Crouch and Leite, 1993; L. Noakes and Paden, 1989).

4.4. *MINIMUM TIME, STRUCTURE INDEPENDENT CONTROL ON THE ESSENTIAL MANIFOLD*

In this section we consider a simple experiment: we want to formulate the control that drives the relative configuration to the desired one in the minimum time, as we have done in section 4.1 for the control on the image plane. We do not make any assumption on the scene, and we want to develop a control strategy which is independent on depth, so that we do not have ill-conditioned controllers at unobservable configurations of the system.

The model described in eq. (11) gives an immediate expression for such a minimum time controller. Suppose we are only interested in maintaining the initial configuration, then $\mathbf{Q}(t_0) = 0$, and our control can be inferred by

$$\mathbf{Q}(t+1|t) = \mathbf{Q}(t|t)\tilde{+}d(t) \tag{21}$$

$$\mathbf{x}^i(t+1)^T(\hat{\mathbf{Q}}(t+1|t) + n(t))\mathbf{x}_0^i = 0 \tag{22}$$

$$\Phi^{-1}(u(t+1)) \doteq -\tilde{+}\hat{\mathbf{Q}}(t+1|t) \tag{23}$$

$$\mathbf{Q}(t+1|t+1) = \mathbf{Q}(t+1|t)\tilde{+} \tilde{+}\Phi^{-1}(u(t+1)) = n(t) \tag{24}$$

and therefore, provided that our estimator is unbiased, the control

$$u(t) = -\Phi(\hat{\mathbf{Q}}(t|t-1))). \tag{25}$$

gives a one-step correction which brings the state to the goal instantaneously up to white, zero-mean noise.

5. EXPERIMENTAL ASSESSMENT

In this section we report an experiment of motion estimation on a real image sequence, and simulations of simple control laws for maintaining a given relative configuration between a scene and an actuated platform on which the camera is mounted. We use as measurements the output of a feature tracking scheme which is a multiscale version of the algorithm developed in (Lucas and Kanade, 1981) implemented on a parallel DSP architecture (TI C40).

The first experiment is described in figure 3 and consists of a box with a checkerboard pattern rotating on top of a chair. The features of the background are rejected by the filter as outliers, and the motion estimates for the feature points which move coherently

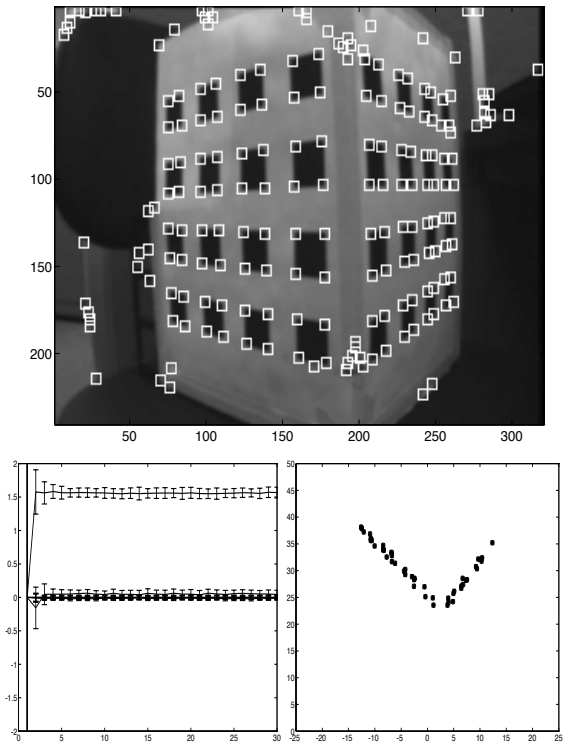

Fig. 3. *The* **"box experiment"***: (top) one frame of the original sequence with the feature points highlighted. (Lower-left) six components of the estimated motion (vertical units are rad/s for the rotational velocity and cm/s for the translational velocity, the horizontal axis is the frame number). (Lower-right) reconstructed scene viewed from the top (the horizontal axis is a slice of the image plane, and the vertical axis is the depth of each feature point).*

with the box are reported in figure 3, as well as the reconstruction of the scene viewed from the top. We used the size of a square of the checkerboard pattern as scale information at the first step.

In the second experiment, described in figure 4, we have simulated a rigid cloud of points moving in front of the camera, which is mounted on some actuated platform, and have generated a simple minimum-time control, based on the motion estimated by the essential filter, in order to maintain the initial configuration between the camera and the scene. The following experiment, reported in figure 5, describes a similar experiment for a different motion of the scene.

In the last experiment, reported in figure 6, we have implemented a minimum-time image-plane control designed for the same task of the previous experiment. In this case the controller is asked to maintain the initial configuration of the points observed on the image plane. Therefore, at each step the controller drives the disparity (difference between projections of the same point at subsequent times, also called visual parallax) to zero. However, we have seen that the image-plane minimum time controller *depends on the depth* (structure) of each point of the scene. When the *structure is known*, then the controller performs similarly to the

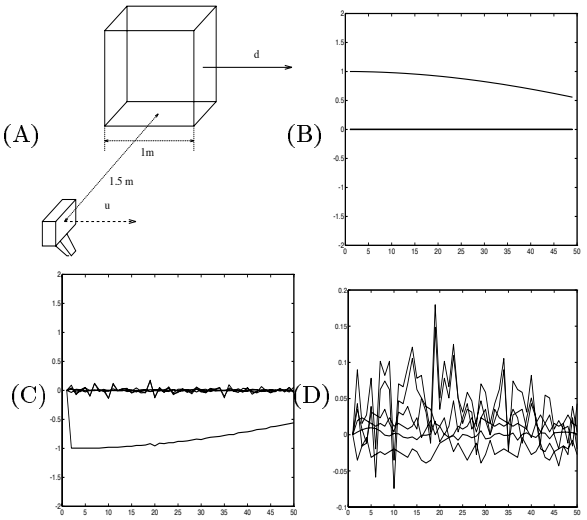

Fig. 4. *"Configuration tracking experiment on the essential space:* **pure translation***": (A) a synthetic scene composed of 30 feature points translates with decreasing translational velocity, the components of which are plotted in (B) in m/s. The minimum time control, whose components are plotted in (C) in m/s, is obtained by feedback from the instantaneous estimate of the relative configuration between the scene and the camera, and quantized at 8 bits. The noise in the image plane was additive white Gaussian with standard deviation corresponding to 10 pixels. The actuators are controlled as to maintain the initial relative configuration between the viewer and the scene; the six local coordinates of the error from the desired configuration are plotted in figure (D) (units are m/s for the error in translational velocity and rad/s for the error in rotational velocity).*

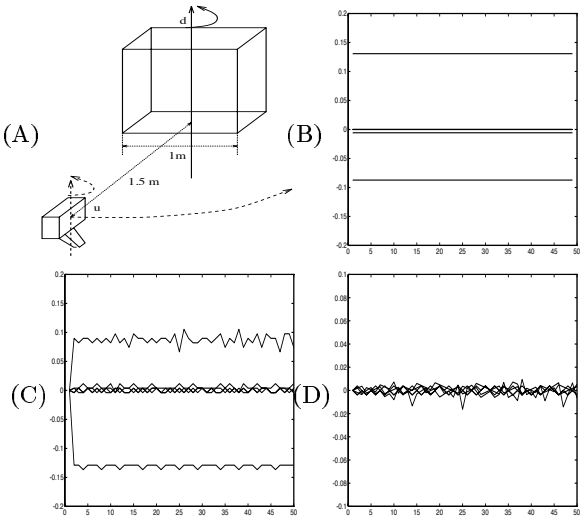

Fig. 5. *"Configuration tracking on the essential space:* **rototraslatory motion***" (A) the scene rotates about a fixed axis which is 1.5m ahead of the observer with constant angular velocity of 5 deg/s. The local coordinates of the relative motion between the scene and the viewer in the viewer's reference are plotted in (B) (m/s for the translational velocity, rad/s for the rotational velocity). The components of the minimum time control are plotted in (C) with the same units, and the corresponding deviation from the desired configuration is plotted in (D). The noise was white, zero-mean and Gaussian with 5 pixel std, and the controller was quantized at 8 bits.*

one on the essential space (see figure 6 (A)-(B)). If the geometry of the scene is not known, then it must be estimated. However, depth cannot be estimated for zero parallax (Soatto, 1994). Therefore the system is affected by the intrinsic conflict between *trying to drive the parallax to zero, and at the same time trying to keep it large enough* in order to be able to compute depth. The effect, which is visible in figure 6 (C)-(D), is that the controller "drifts" in order to accumulate a residual which is large enough for computing depth. In some cases the controller "swings": when the residual is large, there is enough parallax for computing the controller accurately and drive the residual to zero; at this point the controller is computed with gross errors, and the residual grows again.

6. CONCLUSIONS

The purpose of this paper is to stress the flexibility, robustness and accuracy reached by vision as a sensor, which could be considered in alternative to traditional accelerometers or range sensors in a number of applications in robotics. We have shown the image-plane control is not practical when the structure (depth) of the scene is not known, and proposed to perform visual motion control on the essential manifold, using the output of a causal motion estimator, called the "essential filter".

ACKNOWLEDGEMENTS

We would like to thank Prof. Giorgio Picci and Prof. Ruggero Frezza for their advice and support, Prof. Richard Murray for his suggestions and useful comments, Prof. Joel Burdick for his interpretations of the essential space, Francesco Bullo for his comments about the choice of a metric on the essential space and F. Christof Kolb for his criticism about the style of the paper.

This work has been funded by the California Institute of Technology, a scholarship from the University of Padova, a fellowship from the "A. Gini" Foundation and grant ASI-RS-103 from the Italian Space Agency.

7. REFERENCES

Azarbayejani, A., B. Horowitz and A. Pentland (New York, 1993). Recursive estimation of structure and motion using relative orientation constraints. *Proc. CVPR.*

Barron, J., D. Fleet and S. Beauchemin (1992). Performance of optical flow techniques. RPL-TR 9107. Queen's University Kingston, Ontario. Robotics and

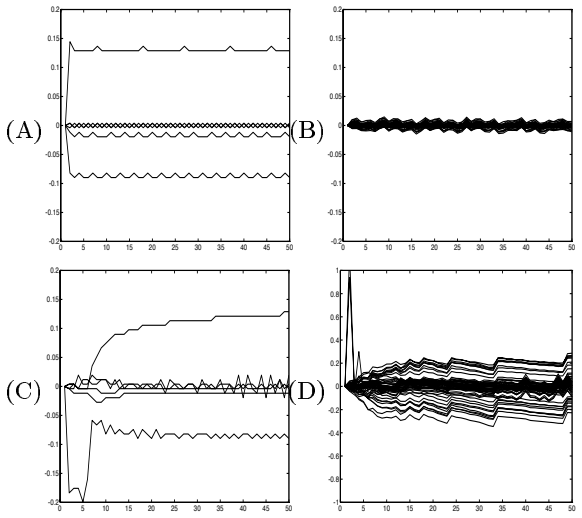

Fig. 6. *"Configuration tracking* **on the image plane***": (A)-(B) for the same experiment described in figure 5, the control on the image plane* when the structure of the scene is known *(in terms of depth of each point) is comparable with the one obtained with the control on the essential manifold, which* does not need information about the structure of the scene *(compare with figure 5 (C)-(D)). When the* structure of the scene is not known, *and depth has to be estimated, the control is far less robust, for it tries to drive the system to a zero-disparity configuration which is ill-conditioned (C)-(D). The controller, whose state depends on the depth of the points in the scene, tries to reduce the image parallax (disparity, or residual) to zero: such configuration, however, does not allow estimating depth. The effect, which is visible in figures (C)-(D), is that the controller "drifts" in order to accumulate a residual which is large enough for computing depth.*

perception laboratory. Also in Proc. CVPR 1992, pp 236-242.

Chaumette, F. and A. Santos (1993). Tracking a moving object by visual servoing. *Proc. of the 12th IFAC World Congr. Vol. 9 pp. 409–414.*

Chevallier, D. P. (1991). Lie algebras, modules, dual quaternions and algebraic methods in kinematics. *Mech. Mach. Theory.*

Crouch, P. and F. Silva Leite (1993). The dynamic interpolation problem on riemannian manifolds, lie groups and symmetric spaces. *Technical report.*

Curwen, R., A. Blake and A. Zisserman (1992). Real-time visual tracking for surveillance and path planning. *Proc. of the ECCV.*

Dickmanns, E. D. and V. Graefe. (1988). Dynamic monocular machine vision. *Machine Vision and Applications* **1**, 223–240.

Ebine, K. Hashimoto T. Kimoto T. and H. Kimura (1991). Image-based dynamic visual servo for a hand-eye manipulator. *Kodama Kimura, editor,* Recent advances in mathematical theory of systems, control, networks, and signal processing II, *pages 609–614. Proceedings of the international symposium of MTNS, Mita Press.*

Faugeras, O. (1993). *Three dimensional vision, a geometric viewpoint.* MIT Press.

Ghosh, B., M Jankovic and Y. Wu (1994). Perspective problems in systems theory and its application in machine vision. *Journal of Math. Systems, Est. and Control.*

Heel, J. (March 1991). Temporal integration of 3-d surface reconstruction. *To appear on IEEE trans. PAMI, special issue on the interpretation of 3-D scenes.*

Ho, C.C. and N.H. McClamroch (1992). Autonomous spacecraft docking using a computer vision system. *Proc. of the 31st CDC Tucson AZ.*

Jazwinski, A.H. (1970). *Stochastic Processes and Filtering Theory.* Academic Press.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Trans. of the ASME-Journal of basic engineering.*

Krener, A. J. and A. Isidori (1983). Linearization by output injection and nonlinear observers. *Systems and Control Letters vol. 3.*

L. Noakes, G. Heinzinger and B. Paden (1989). Cubic splines on curved spaces. *IMA J. Math. Control and Information.*

Lei, Ming and Bijoy K. Ghosh (1993). A new nonlinear feedback controller for visually-guided robotic motion tracking. *Proc. of the ECC.*

Longuet-Higgins, H. C. (1981). A computer algorithm for reconstructing a scene from two projections. *Nature* **293**, 133–135.

Lucas, B.D. and T. Kanade (1981). An iterative image registration technique with an application to stereo vision.. *Proc. 7th Int. Joint Conf. on Art. Intell.*

Matthies, L., R. Szeliski and T. Kanade (1989). Kalman filter-based algorithms for estimating depth from image sequences. *Int. J. of computer vision.*

Maybank, S. (1992). *Theory of reconstruction from image motion.* Springer Verlag.

Murray, R.M., Z. Li and S.S. Sastry (1993). *A Mathematical Introduction to Robotic Manipulation.* CRC Press.

Oliensis, J. and J. Inigo-Thomas (1992). Recursive multi-frame structure from motion incorporating motion error. *Proc. DARPA Image Understanding Workshop.*

Soatto, S. (1994). Observability/identifiability of rigid motion under perspective. *Technical Report CIT-CDS 94-001, California Institute of Technology (available through the WWW net MOSAIC http://avalon.caltech.edu/cds/techreports/). Submitted to "Automatica".*

Soatto, S. and P. Perona (1994*a*). On the exact linearization of structure from motion. *Technical Report CIT-CDS 94-011, California Institute of Technology (available through the WWW net MOSAIC http://avalon.caltech.edu/cds/techreports). Submitted to the Int. J. of Computer Vision.*

Soatto, S., P. Perona, R. Frezza and G. Picci (1993). Recursive motion and structure estimation with complete error characterization. In: *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*. New York. pp. 428–433.

Soatto, S., R. Frezza and P. Perona (1994*b*). Motion estimation via dynamic vision. *Submitted to the IEEE trans. on Automatic Control. Registered as Technical Report CIT-CDS-94-004, California Institute of Technology (available through the WWW MOSAIC http://avalon.caltech.edu/cds/techreports).*

Soatto, S., R. Frezza and P. Perona (May 1994*a*). Motion estimation on the essential manifold. *In "Computer Vision ECCV 94, Lecture Notes in Computer Sciences vol. 801", Springer Verlag.*